Exploiting Environment Configurability in Reinforcement Learning

Alberto Maria Metelli

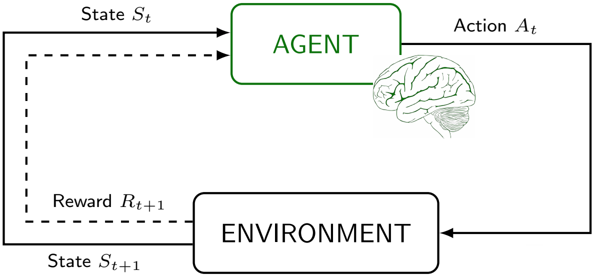
albertomaria.metelli@polimi.it

Supervisor: Prof. Marcello Restelli
Tutor: Prof. Nicola Gatti

Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria
Doctoral Programme in Information Technology - Cycle XXXIII

11th March 2021

Reinforcement Learning

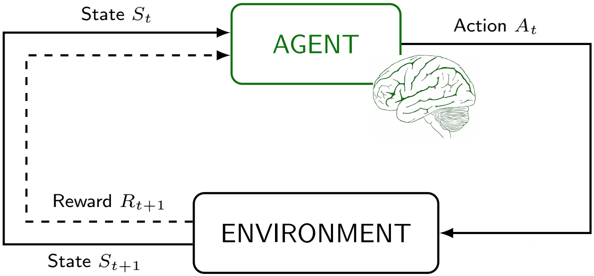


- Markov Decision Process (MDP, Puterman, 2014)
  1. Observe the state S_t
  2. Perform an action $A_t \sim \pi(\cdot|S_t)$
  3. Transition to the next state $S_{t+1} \sim P(\cdot|S_t, A_t)$
  4. Obtain reward $R_{t+1} = r(S_t, A_t, S_{t+1})$

- **Goal**: maximize the expected cumulative discounted reward (Sutton and Barto, 2018):

$$\pi^* \in \arg\max_{\pi \in \Pi^{SR}} J^{\pi} = \mathbb{E}^{\pi}\left[\sum_{t \in \mathbb{N}} \gamma^t R_{t+1}\right]$$

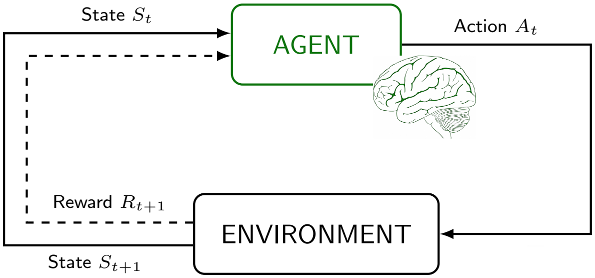
Reinforcement Learning

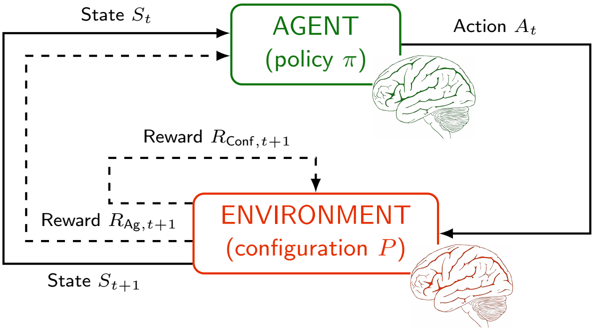


- Markov Decision Process (MDP, Puterman, 2014)
  1. Observe the state S_t
  2. Perform an action $A_t \sim \pi(\cdot|S_t)$
  3. Transition to the next state $S_{t+1} \sim P(\cdot|S_t, A_t)$
  4. Obtain reward $R_{t+1} = r(S_t, A_t, S_{t+1})$

- **Goal**: maximize the expected cumulative discounted reward (Sutton and Barto, 2018):

$$\pi^* \in \arg\max_{\pi \in \Pi^{SR}} J^{\pi} = \mathbb{E}^{\pi}\left[\sum_{t \in \mathbb{N}} \gamma^t R_{t+1}\right]$$


State S_t
AGENT
Action A_t
Reward R_{t+1}
ENVIRONMENT
State S_{t+1}



What if some parts of the environment are **configurable**?

F1 Driving

F1 Driving

- Goal of the configuration:
  - Find the configuration **best suited** for the agent
  - Present different configurations to **speed up** learning

F1 Driving

- Goal of the configuration:
  - Find the configuration **best suited** for the agent
  - Present different configurations to **speed up** learning
- Configuration carried out by **agent** or **external configurator**

F1 Driving

- Goal of the configuration:
  - Find the configuration **best suited** for the agent
  - Present different configurations to **speed up** learning
- Configuration carried out by **agent** or **external configurator**
- Same goal for agent and configurator: **cooperative** setting

Teacher-Student

Teacher-Student

- Again **cooperative** setting

Teacher-Student

- Again **cooperative** setting
- Configuration activity aware of the **agent's capabilities**

Teacher-Student

- Again **cooperative** setting
- Configuration activity aware of the **agent's capabilities**
- Side goal: **infer** the agent's capabilities

Supermarket

Supermarket

- Agent and configurator with different goals: **non-cooperative** setting

Supermarket

- Agent and configurator with different goals: **non-cooperative** setting
- Different modes of interactions:
  - Agent is **aware** of the configurator
  - Agent is **unaware** of the configurator

Outline of the Contributions

I - Modeling Environment Configurability

Configurable Markov Decision Process
(Metelli at al., 2018a, ICML)

Cooperative vs Non-Cooperative
(Ramponi at al., 2021a, AAAI workshop)

II - Learning in cooperative Conf-MDPs

Finite and known environments
(Metelli at al., 2018a, ICML)

Continuous and unknown environments
(Metelli et al., 2019a, ICML)

III - Applications of Conf-MDPs

Policy Space Identification
(Metelli et al. 2019b, under revision MLJ)

Control Frequency Adaptation
(Metelli et al., 2020, ICML)

Outline of the Contributions

I - Modeling Environment Configurability

Configurable Markov Decision Process
(Metelli at al., 2018a, ICML)

Cooperative vs Non-Cooperative
(Ramponi at al., 2021a, AAAI workshop)

II - Learning in cooperative Conf-MDPs

Finite and known environments
(Metelli at al., 2018a, ICML)

Continuous and unknown environments
(Metelli et al., 2019a, ICML)

III - Applications of Conf-MDPs

Policy Space Identification
(Metelli et al. 2019b, under revision MLJ)

Control Frequency Adaptation
(Metelli et al., 2020, ICML)

Outline of the Contributions

I - Modeling Environment Configurability

Configurable Markov Decision Process
(Metelli at al., 2018a, ICML)

Cooperative vs Non-Cooperative
(Ramponi at al., 2021a, AAAI workshop)

II - Learning in cooperative Conf-MDPs

Finite and known environments
(Metelli at al., 2018a, ICML)

Continuous and unknown environments
(Metelli et al., 2019a, ICML)

III - Applications of Conf-MDPs

Policy Space Identification
(Metelli et al. 2019b, under revision MLJ)

Control Frequency Adaptation
(Metelli et al., 2020, ICML)

Part I - Modeling Environment Configurability

I - Modeling Environment Configurability

Configurable Markov Decision Process
(Metelli at al., 2018a, ICML)

Cooperative vs Non-Cooperative
(Ramponi at al., 2021a, AAAI workshop)

II - Learning in cooperative Conf-MDPs

Finite and known environments
(Metelli at al., 2018a, ICML)

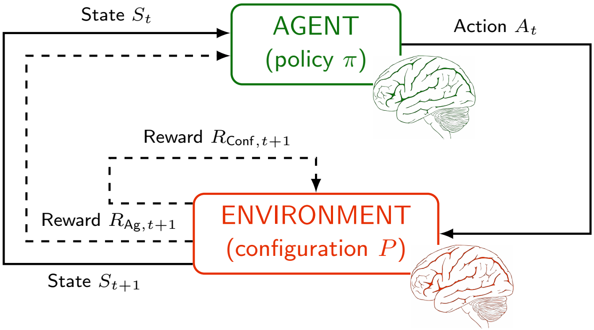
Continuous and unknown environments
(Metelli et al., 2019a, ICML)

III - Applications of Conf-MDPs

Policy Space Identification
(Metelli et al. 2019b, under revision MLJ)

Control Frequency Adaptation
(Metelli et al., 2020, ICML)

Reinforcement Learning in Configurable Environments



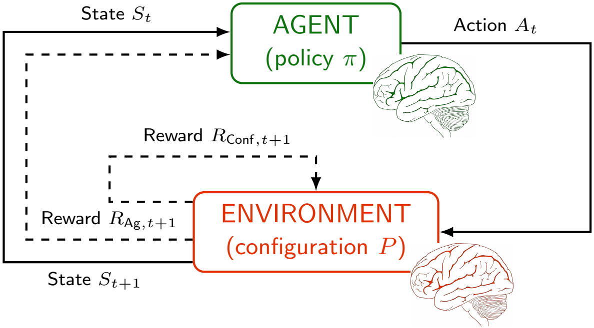
- **Configurable** Markov Decision Process (Conf-MDP)
  1. Observe the state S_t
  2. Perform an action $A_t \sim \pi(\cdot|S_t)$
  3. Transition to the next state $S_{t+1} \sim P(\cdot|S_t, A_t)$
  4. Agent obtains reward $R_{t+1,Ag} = r_{Ag}(S_t, A_t, S_{t+1})$
  5. Configurator obtains reward $R_{t+1,Conf} = r_{Conf}(S_t, A_t, S_{t+1})$

- Expected cumulative discounted reward for agent and configurator:

$$J_{Ag}^{\pi,P} = \mathbb{E}^{\pi,P}\left[\sum_{t\in\mathbb{N}} \gamma^t R_{Ag,t+1}\right] \qquad J_{Conf}^{\pi,P} = \mathbb{E}^{\pi,P}\left[\sum_{t\in\mathbb{N}} \gamma^t R_{Conf,t+1}\right]$$

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Reinforcement Learning in Configurable Environments



- **Configurable** Markov Decision Process (Conf-MDP)
  1. Observe the state S_t
  2. Perform an action $A_t \sim \pi(\cdot|S_t)$
  3. Transition to the next state $S_{t+1} \sim P(\cdot|S_t, A_t)$
  4. Agent obtains reward $R_{t+1,Ag} = r_{Ag}(S_t, A_t, S_{t+1})$
  5. Configurator obtains reward $R_{t+1,Conf} = r_{Conf}(S_t, A_t, S_{t+1})$

- Expected cumulative discounted reward for agent and configurator:

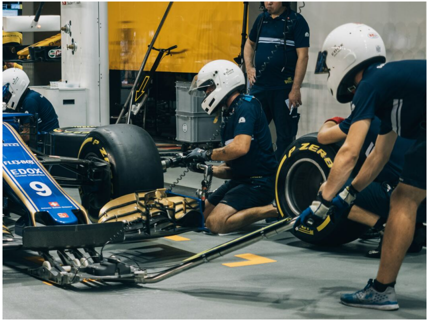
$$J^{\pi,P}_{Ag} = \mathbb{E}^{\pi,P}\left[\sum_{t\in\mathbb{N}} \gamma^t R_{Ag,t+1}\right] \qquad J^{\pi,P}_{Conf} = \mathbb{E}^{\pi,P}\left[\sum_{t\in\mathbb{N}} \gamma^t R_{Conf,t+1}\right]$$

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Cooperative and Non-Cooperative Settings

Cooperative Conf-MDP

$$r_{Ag} = r_{Conf} =: r$$



Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Non-Cooperative Conf-MDP

$$r_{Ag} \neq r_{Conf}$$

Giorgia Ramponi, Alberto Maria Metelli, Alessandro Concetti, and Marcello Restelli. *Online Learning in Non-Cooperative Configurable Markov Decision Process*. AAAI-21 Workshop on Reinforcement Learning in Games, 2021.

Cooperative and Non-Cooperative Settings

Cooperative Conf-MDP

$$r_{Ag} = r_{Conf} =: r$$

$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi, P}$$

- Simple definition of *optimality*
- Π and $\mathcal{P}$ policy and configuration spaces

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Non-Cooperative Conf-MDP

$$r_{Ag} \neq r_{Conf}$$

Giorgia Ramponi, Alberto Maria Metelli, Alessandro Concetti, and Marcello Restelli. *Online Learning in Non-Cooperative Configurable Markov Decision Process*. AAAI-21 Workshop on Reinforcement Learning in Games, 2021.

Cooperative and Non-Cooperative Settings

Cooperative Conf-MDP

$$r_{Ag} = r_{Conf} =: r$$

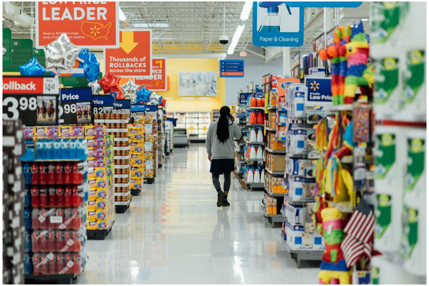
$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi, P}$$

- Simple definition of *optimality*
- Π and $\mathcal{P}$ policy and configuration spaces

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Non-Cooperative Conf-MDP

$$r_{Ag} \neq r_{Conf}$$



Giorgia Ramponi, Alberto Maria Metelli, Alessandro Concetti, and Marcello Restelli. *Online Learning in Non-Cooperative Configurable Markov Decision Process*. AAAI-21 Workshop on Reinforcement Learning in Games, 2021.

Cooperative and Non-Cooperative Settings

Cooperative Conf-MDP

$$r_{Ag} = r_{Conf} =: r$$

$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi, P}$$

- Simple definition of *optimality*
- Π and $\mathcal{P}$ policy and configuration spaces

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Non-Cooperative Conf-MDP

$$r_{Ag} \neq r_{Conf}$$

$$P^* \in \arg\max_{P \in \mathcal{P}} J_{Conf}^{\pi^{BR(P)}, P}$$

$$\pi^{BR(P)} \in \arg\max_{\pi \in \Pi} J_{Ag}^{\pi, P}$$

- Equilibria as solution concepts (e.g., *Stackelberg* (Von Stackelberg, 1934))
- To be further studied...

Giorgia Ramponi, Alberto Maria Metelli, Alessandro Concetti, and Marcello Restelli. *Online Learning in Non-Cooperative Configurable Markov Decision Process*. AAAI-21 Workshop on Reinforcement Learning in Games, 2021.

Considerations

- Configuration **limited** to a portion of the environment → **parametric** setting

$$P_\omega \in \mathcal{P}$$

- Configuration happens **less frequently** than policy update and might be **expensive** (Silva et al., 2018)

$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi, P} - \mathrm{Cost}(P)$$

- Solving a cooperative Conf-MDP for general configuration space $\mathcal{P}$ is **NP-Hard** (Silva et al., 2019)

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Considerations

- Configuration **limited** to a portion of the environment → **parametric** setting

$$P_\omega \in \mathcal{P}$$

- Configuration happens **less frequently** than policy update and might be **expensive** (Silva et al., 2018)

$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi,P} - \text{Cost}(P)$$

- Solving a cooperative Conf-MDP for general configuration space $\mathcal{P}$ is **NP-Hard** (Silva et al., 2019)

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Considerations

- Configuration **limited** to a portion of the environment → **parametric** setting

$$P_\omega \in \mathcal{P}$$

- Configuration happens **less frequently** than policy update and might be **expensive** (Silva et al., 2018)

$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi, P} - \text{Cost}(P)$$

- Solving a cooperative Conf-MDP for general configuration space $\mathcal{P}$ is **NP-Hard** (Silva et al., 2019)

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Part II - Learning in cooperative Conf-MDPs

I - Modeling Environment Configurability

Configurable Markov Decision Process
(Metelli at al., 2018a, ICML)

Cooperative vs Non-Cooperative
(Ramponi at al., 2021a, AAAI workshop)

II - Learning in cooperative Conf-MDPs

Finite and known environments
(Metelli at al., 2018a, ICML)

Continuous and unknown environments
(Metelli et al., 2019a, ICML)

III - Applications of Conf-MDPs

Policy Space Identification
(Metelli et al. 2019b, under revision MLJ)

Control Frequency Adaptation
(Metelli et al., 2020, ICML)

Learning Algorithms for Cooperative Conf-MDPs

$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi, P}$$

Safe Policy Model Iteration (SPMI)

- **Finite** state-action spaces
- **Known** configuration space $\mathcal{P}$
- **Monotonic** performance improvement (Kakade and Langford, 2002)

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Relative Entropy Model Policy Search (REMPS)

- **Trust-region** method (Peters et al., 2010)
- **Continuous** state-action spaces
- **Learned** configuration space $\widehat{\mathcal{P}}$ from data

Alberto Maria Metelli, Emanuele Ghelfi, and Marcello Restelli. *Reinforcement Learning in Configurable Continuous Environments*. Proceedings of the 36th International Conference on Machine Learning, ICML 2019.

Learning Algorithms for Cooperative Conf-MDPs

$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi, P}$$

Safe Policy Model Iteration (SPMI)

- **Finite** state-action spaces
- **Known** configuration space $\mathcal{P}$
- **Monotonic** performance improvement (Kakade and Langford, 2002)

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

Relative Entropy Model Policy Search (REMPS)

- **Trust-region** method (Peters et al., 2010)
- **Continuous** state-action spaces
- **Learned** configuration space $\widehat{\mathcal{P}}$ from data

Alberto Maria Metelli, Emanuele Ghelfi, and Marcello Restelli. *Reinforcement Learning in Configurable Continuous Environments*. Proceedings of the 36th International Conference on Machine Learning, ICML 2019.

Learning Algorithms for Cooperative Conf-MDPs

$$\pi^*, P^* \in \arg\max_{\pi \in \Pi, P \in \mathcal{P}} J^{\pi, P}$$

Safe Policy Model Iteration (SPMI)

- **Finite** state-action spaces
- **Known** configuration space $\mathcal{P}$
- **Monotonic** performance improvement (Kakade and Langford, 2002)

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. *Configurable Markov Decision Processes*. Proceedings of the 35th International Conference on Machine Learning, ICML 2018.

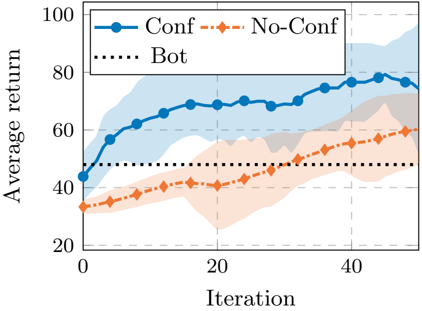
Relative Entropy Model Policy Search (REMPS)

- **Trust-region** method (Peters et al., 2010)
- **Continuous** state-action spaces
- **Learned** configuration space $\widehat{\mathcal{P}}$ from data

Alberto Maria Metelli, Emanuele Ghelfi, and Marcello Restelli. *Reinforcement Learning in Configurable Continuous Environments*. Proceedings of the 36th International Conference on Machine Learning, ICML 2019.

Learning to Configure Vehicle with TORCS

- **Policy**: acceleration, steer, brake (Wymann et al., 2000)
- **Configurable Parameters**
  - rear wing angle
  - front wing angle
  - brake repartition



Alberto Maria Metelli, Emanuele Ghelfi, and Marcello Restelli. *Reinforcement Learning in Configurable Continuous Environments*. Proceedings of the 36th International Conference on Machine Learning, ICML 2019.

Part III - Applications of Conf-MDPs

I - Modeling Environment Configurability

Configurable Markov Decision Process
(Metelli at al., 2018a, ICML)

Cooperative vs Non-Cooperative
(Ramponi at al., 2021a, AAAI workshop)

II - Learning in cooperative Conf-MDPs

Finite and known environments
(Metelli at al., 2018a, ICML)

Continuous and unknown environments
(Metelli et al., 2019a, ICML)

III - Applications of Conf-MDPs

Policy Space Identification
(Metelli et al. 2019b, under revision MLJ)

Control Frequency Adaptation
(Metelli et al., 2020, ICML)

Policy Space Identification

I - Modeling Environment Configurability

Configurable Markov Decision Process
(Metelli at al., 2018a, ICML)

Cooperative vs Non-Cooperative
(Ramponi at al., 2021a, AAAI workshop)

II - Learning in cooperative Conf-MDPs

Finite and known environments
(Metelli at al., 2018a, ICML)

Continuous and unknown environments
(Metelli et al., 2019a, ICML)

III - Applications of Conf-MDPs

Policy Space Identification
(Metelli et al. 2019b, under revision MLJ)

Control Frequency Adaptation
(Metelli et al., 2020, ICML)

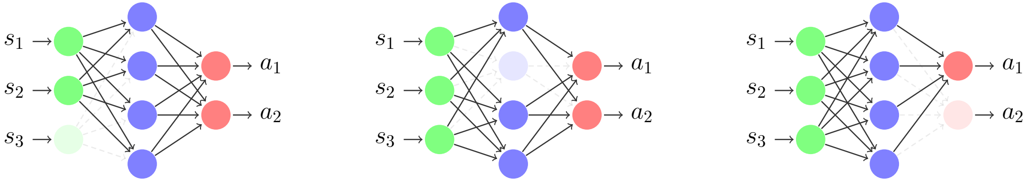
Motivations and Problem

- **Problem**: The configurator should know the **perception** and **actuation** capabilities of an agent to select a suitable configuration
- **Research Question**: How to identify the **policy space** of an agent by observing its behavior?
- Applications
  - Configurable MDPs
  - Imitation Learning (Osa et al., 2018)

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Motivations and Problem

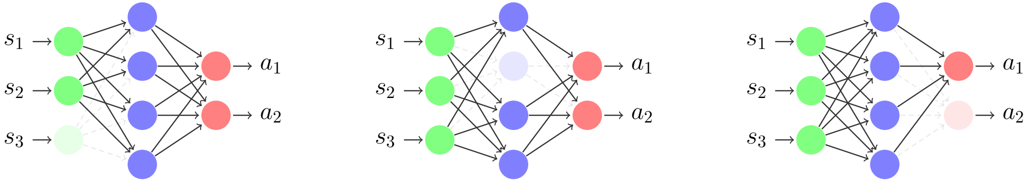
- **Problem**: The configurator should know the **perception** and **actuation** capabilities of an agent to select a suitable configuration
- **Research Question**: How to identify the **policy space** of an agent by observing its behavior?
- Applications
  - Configurable MDPs
  - Imitation Learning (Osa et al., 2018)



Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Motivations and Problem

- **Problem**: The configurator should know the **perception** and **actuation** capabilities of an agent to select a suitable configuration
- **Research Question**: How to identify the **policy space** of an agent by observing its behavior?
- Applications
  - Configurable MDPs
  - Imitation Learning (Osa et al., 2018)



Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

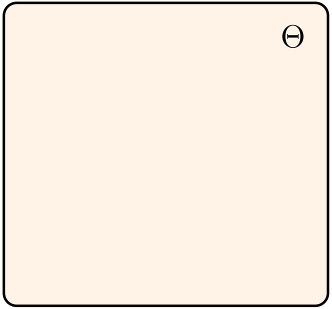
Policy Spaces and Correctness

- Agent policy $\rightarrow$ $\pi_{\boldsymbol{\theta}^*} \in \Pi_{\Theta}$ $\leftarrow$ Policy space
- Parameter space $\Theta \subset \mathbb{R}^d$
- The agent can change $d^* < d$ parameters
- $I \subseteq \{1, \dots, d\}$ subset of indexes

$$\Theta_I = \{\boldsymbol{\theta} \in \Theta : \theta_i = 0, \forall i \in \{1, \dots, d\} \backslash I\}$$

- I^* is **correct** for the agent's policy $\pi_{\boldsymbol{\theta}^*}$ iff

$$\underbrace{\boldsymbol{\theta}^* \in \Theta_{I^*}}_{\text{sufficient}} \quad \wedge \quad \underbrace{\forall i \in I^* : \boldsymbol{\theta}^* \notin \Theta_{I^* \backslash \{i\}}}_{\text{necessary}}$$



Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

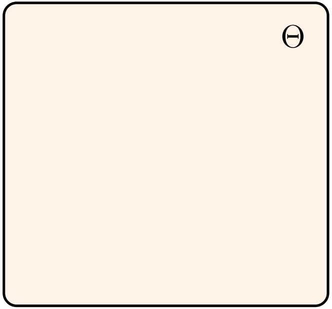
Policy Spaces and Correctness

- Agent policy $\rightarrow$ $\pi_{\boldsymbol{\theta}^*} \in \Pi_{\Theta}$ $\leftarrow$ Policy space
- Parameter space $\Theta \subset \mathbb{R}^d$
- The agent can change $d^* < d$ parameters
- $I \subseteq \{1, \dots, d\}$ subset of indexes

$$\Theta_I = \{\boldsymbol{\theta} \in \Theta : \theta_i = 0, \forall i \in \{1, \dots, d\} \setminus I\}$$

- I^* is **correct** for the agent's policy $\pi_{\boldsymbol{\theta}^*}$ iff

$$\underbrace{\boldsymbol{\theta}^* \in \Theta_{I^*}}_{\text{sufficient}} \quad \wedge \quad \underbrace{\forall i \in I^* : \boldsymbol{\theta}^* \notin \Theta_{I^* \setminus \{i\}}}_{\text{necessary}}$$



Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

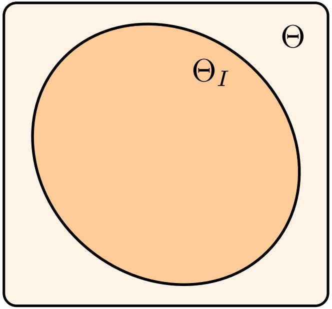
Policy Spaces and Correctness

- Agent policy $\rightarrow$ $\pi_{\boldsymbol{\theta}^*} \in \Pi_\Theta$ $\leftarrow$ Policy space
- Parameter space $\Theta \subset \mathbb{R}^d$
- The agent can change $d^* < d$ parameters
- $I \subseteq \{1, \dots, d\}$ subset of indexes

$$\Theta_I = \{\boldsymbol{\theta} \in \Theta : \theta_i = 0, \forall i \in \{1, \dots, d\} \setminus I\}$$

- I^* is **correct** for the agent's policy $\pi_{\boldsymbol{\theta}^*}$ iff

$$\underbrace{\boldsymbol{\theta}^* \in \Theta_{I^*}}_{\text{sufficient}} \quad \wedge \quad \underbrace{\forall i \in I^* : \boldsymbol{\theta}^* \notin \Theta_{I^* \setminus \{i\}}}_{\text{necessary}}$$



<u>Alberto Maria Metelli</u>, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

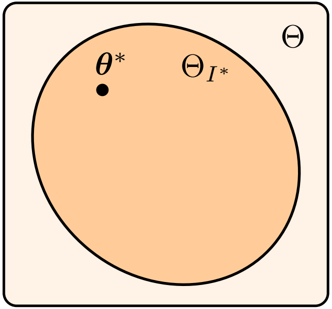
Policy Spaces and Correctness

- Agent policy $\longrightarrow$ $\pi_{\boldsymbol{\theta}^*} \in \Pi_\Theta$ $\longleftarrow$ Policy space
- Parameter space $\Theta \subset \mathbb{R}^d$
- The agent can change $d^* < d$ parameters
- $I \subseteq \{1, \dots, d\}$ subset of indexes

$$\Theta_I = \{\boldsymbol{\theta} \in \Theta : \theta_i = 0, \forall i \in \{1, \dots, d\} \setminus I\}$$

- I^* is **correct** for the agent's policy $\pi_{\boldsymbol{\theta}^*}$ iff

$$\underbrace{\boldsymbol{\theta}^* \in \Theta_{I^*}}_{\text{sufficient}} \quad \wedge \quad \underbrace{\forall i \in I^* : \boldsymbol{\theta}^* \notin \Theta_{I^* \setminus \{i\}}}_{\text{necessary}}$$



Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

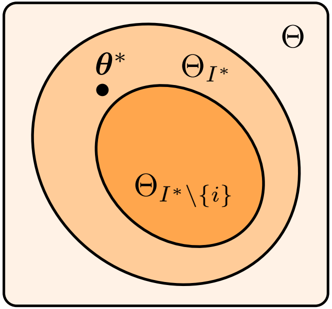
Policy Spaces and Correctness

- Agent policy $\rightarrow$ $\pi_{\boldsymbol{\theta}^*} \in \Pi_\Theta$ $\leftarrow$ Policy space
- Parameter space $\Theta \subset \mathbb{R}^d$
- The agent can change $d^* < d$ parameters
- $I \subseteq \{1, \dots, d\}$ subset of indexes

$$\Theta_I = \{\boldsymbol{\theta} \in \Theta : \theta_i = 0, \forall i \in \{1, \dots, d\} \backslash I\}$$

- I^* is **correct** for the agent's policy $\pi_{\boldsymbol{\theta}^*}$ iff

$$\underbrace{\boldsymbol{\theta}^* \in \Theta_{I^*}}_{\textbf{sufficient}} \quad \wedge \quad \underbrace{\forall i \in I^* : \boldsymbol{\theta}^* \notin \Theta_{I^* \backslash \{i\}}}_{\textbf{necessary}}$$



Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Hypothesis Tests

- **Idea**: perform **hypothesis test** for $I \subseteq \{1, \dots, d\}$

$$\mathcal{H}_{0,I} : \boldsymbol{\theta}^* \in \Theta_I \quad \text{vs} \quad \mathcal{H}_{1,I} : \boldsymbol{\theta}^* \in \Theta \backslash \Theta_I$$

- Dataset of samples $\{(S_i, A_i)\}_{i=1}^n$ collected with the agent's policy $\pi_{\boldsymbol{\theta}^*}$
- Likelihood of a parameter $\boldsymbol{\theta} \in \Theta$

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{\boldsymbol{\theta}}(A_i | S_i)$$

- Generalized **likelihood ratio** statistic (Casella and Berger, 2002)

$$\Lambda_I = \frac{\sup_{\boldsymbol{\theta} \in \Theta_I} \widehat{\mathcal{L}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})}$$

$\Lambda_I \simeq 0 \rightarrow$ reject $\mathcal{H}_{0,I}$

$\Lambda_I \simeq 1 \rightarrow$ do not reject $\mathcal{H}_{0,I}$

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Hypothesis Tests

- **Idea**: perform **hypothesis test** for $I \subseteq \{1, \dots, d\}$

$$\mathcal{H}_{0,I} : \boldsymbol{\theta}^* \in \Theta_I \quad \text{vs} \quad \mathcal{H}_{1,I} : \boldsymbol{\theta}^* \in \Theta \backslash \Theta_I$$

- Dataset of samples $\{(S_i, A_i)\}_{i=1}^n$ collected with the agent's policy $\pi_{\boldsymbol{\theta}^*}$
- Likelihood of a parameter $\theta \in \Theta$

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{\boldsymbol{\theta}}(A_i|S_i)$$

- Generalized **likelihood ratio** statistic (Casella and Berger, 2002)

$$\Lambda_I = \frac{\sup_{\boldsymbol{\theta} \in \Theta_I} \widehat{\mathcal{L}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})}$$

$\Lambda_I \simeq 0 \rightarrow$ reject $\mathcal{H}_{0,I}$

$\Lambda_I \simeq 1 \rightarrow$ do not reject $\mathcal{H}_{0,I}$

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Hypothesis Tests

- **Idea**: perform **hypothesis test** for $I \subseteq \{1, \dots, d\}$

$$\mathcal{H}_{0,I} \ : \ \boldsymbol{\theta}^* \in \Theta_I \quad \text{vs} \quad \mathcal{H}_{1,I} \ : \ \boldsymbol{\theta}^* \in \Theta \backslash \Theta_I$$

- Dataset of samples $\{(S_i, A_i)\}_{i=1}^n$ collected with the agent's policy $\pi_{\boldsymbol{\theta}^*}$
- Likelihood of a parameter $\boldsymbol{\theta} \in \Theta$

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{\boldsymbol{\theta}}(A_i|S_i)$$

- Generalized **likelihood ratio** statistic (Casella and Berger, 2002)

$$\Lambda_I = \frac{\sup_{\boldsymbol{\theta} \in \Theta_I} \widehat{\mathcal{L}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})}$$

$\Lambda_I \simeq 0 \ \longrightarrow$ reject $\mathcal{H}_{0,I}$
$\Lambda_I \simeq 1 \ \longrightarrow$ do not reject $\mathcal{H}_{0,I}$

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Hypothesis Tests

- **Idea**: perform **hypothesis test** for $I \subseteq \{1, \dots, d\}$

$$\mathcal{H}_{0,I} : \boldsymbol{\theta}^* \in \Theta_I \quad \text{vs} \quad \mathcal{H}_{1,I} : \boldsymbol{\theta}^* \in \Theta \backslash \Theta_I$$

- Dataset of samples $\{(S_i, A_i)\}_{i=1}^n$ collected with the agent's policy $\pi_{\boldsymbol{\theta}^*}$
- Likelihood of a parameter $\boldsymbol{\theta} \in \Theta$

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{\boldsymbol{\theta}}(A_i | S_i)$$

- Generalized **likelihood ratio** statistic (Casella and Berger, 2002)

$$\Lambda_I = \frac{\sup_{\boldsymbol{\theta} \in \Theta_I} \widehat{\mathcal{L}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})}$$

$\Lambda_I \simeq 0 \rightarrow$ reject $\mathcal{H}_{0,I}$
$\Lambda_I \simeq 1 \rightarrow$ do not reject $\mathcal{H}_{0,I}$

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Hypothesis Tests

- **Idea**: perform **hypothesis test** for $I \subseteq \{1, \dots, d\}$

$$\mathcal{H}_{0,I} \;:\; \boldsymbol{\theta}^* \in \Theta_I \quad \text{vs} \quad \mathcal{H}_{1,I} \;:\; \boldsymbol{\theta}^* \in \Theta \backslash \Theta_I$$

- Dataset of samples $\{(S_i, A_i)\}_{i=1}^n$ collected with the agent's policy $\pi_{\boldsymbol{\theta}^*}$
- Likelihood of a parameter $\boldsymbol{\theta} \in \Theta$

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{\boldsymbol{\theta}}(A_i|S_i)$$

- Generalized **likelihood ratio** statistic (Casella and Berger, 2002)

$$\Lambda_I = \frac{\sup_{\boldsymbol{\theta} \in \Theta_I} \widehat{\mathcal{L}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})}$$

$\Lambda_I \simeq 0 \;\rightarrow\;$ reject $\mathcal{H}_{0,I}$
$\Lambda_I \simeq 1 \;\rightarrow\;$ do not reject $\mathcal{H}_{0,I}$

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Identification Rules

- **Identification Rule**: retain all the **approximately correct** $\widehat{I} \subseteq \{1, \dots, d\}$:

$$\underbrace{\text{do not reject } \mathcal{H}_{0,\widehat{I}}}_{\textbf{sufficient}} \quad \wedge \quad \underbrace{\forall i \in \widehat{I} : \text{reject } \mathcal{H}_{0,\widehat{I}\setminus\{i\}}}_{\textbf{necessary}}$$

- Can be simplified under **uniqueness** of representation
- Theoretical guarantees on misidentification

$$\Pr\left(\widehat{I} \neq I^*\right) \leqslant \mathcal{O}\left(d^2 \exp\left(-\frac{c(\theta^*)n}{16d^2\sigma^4}\right)\right)$$

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Identification Rules

- **Identification Rule**: retain all the **approximately correct** $\widehat{I} \subseteq \{1, \dots, d\}$:

$$\underbrace{\text{do not reject } \mathcal{H}_{0,\widehat{I}}}_{\text{sufficient}} \quad \wedge \quad \underbrace{\forall i \in \widehat{I} : \text{reject } \mathcal{H}_{0,\widehat{I} \setminus \{i\}}}_{\text{necessary}}$$

- Can be simplified under **uniqueness** of representation
- Theoretical guarantees on misidentification

$$\Pr\left(\widehat{I} \neq I^*\right) \leqslant \mathcal{O}\left(d^2 \exp\left(-\frac{c(\theta^*)n}{16 d^2 \sigma^4}\right)\right)$$

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Identification Rules

- **Identification Rule**: retain all the **approximately correct** $\widehat{I} \subseteq \{1, \dots, d\}$:

$$\underbrace{\text{do not reject } \mathcal{H}_{0,\widehat{I}}}_{\textbf{sufficient}} \quad \wedge \quad \underbrace{\forall i \in \widehat{I} : \text{reject } \mathcal{H}_{0,\widehat{I} \setminus \{i\}}}_{\textbf{necessary}}$$

- Can be simplified under **uniqueness** of representation
- Theoretical guarantees on misidentification

$$\Pr\left(\widehat{I} \neq I^*\right) \leqslant \mathcal{O}\left(d^2 \exp\left(-\frac{c(\boldsymbol{\theta}^*)n}{16 d^2 \sigma^4}\right)\right)$$

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. *Policy Space Identification in Configurable Environments*. CoRR, abs/1909.03984, 2019b.

Control Frequency Adaptation

I - Modeling Environment Configurability

Configurable Markov Decision Process
(Metelli at al., 2018a, ICML)

Cooperative vs Non-Cooperative
(Ramponi at al., 2021a, AAAI workshop)

II - Learning in cooperative Conf-MDPs

Finite and known environments
(Metelli at al., 2018a, ICML)

Continuous and unknown environments
(Metelli et al., 2019a, ICML)

III - Applications of Conf-MDPs

Policy Space Identification
(Metelli et al. 2019b, under revision MLJ)

Control Frequency Adaptation
(Metelli et al., 2020, ICML)

Motivations and Problem

- **Problem**: The **control frequency** for a system is a **configurable** environmental parameter.
- Applications
  - Robot control (Kober et al., 2013)
  - Finance, trading (Murphy et al., 2001)

	Control opportunities	Sample complexity
High frequency	✓	✗
Low frequency	✗	✓

- **Research Question**: Can we exploit this **trade-off** to find an **optimal** control frequency?

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Motivations and Problem

- **Problem**: The **control frequency** for a system is a **configurable** environmental parameter.
- Applications
  - Robot control (Kober et al., 2013)
  - Finance, trading (Murphy et al., 2001)

	Control opportunities	Sample complexity
High frequency	✓	✗
Low frequency	✗	✓

- **Research Question**: Can we exploit this **trade-off** to find an **optimal** control frequency?

<u>Alberto Maria Metelli</u>, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Motivations and Problem

- **Problem**: The **control frequency** for a system is a **configurable** environmental parameter.
- Applications
  - Robot control (Kober et al., 2013)
  - Finance, trading (Murphy et al., 2001)

	Control opportunities	Sample complexity
High frequency	✓	✗
Low frequency	✗	✓

- **Research Question**: Can we exploit this **trade-off** to find an **optimal** control frequency?

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Motivations and Problem

- **Problem**: The **control frequency** for a system is a **configurable** environmental parameter.
- Applications
  - Robot control (Kober et al., 2013)
  - Finance, trading (Murphy et al., 2001)

	Control opportunities	Sample complexity
High frequency	✓	✗
Low frequency	✗	✓

- **Research Question**: Can we exploit this **trade-off** to find an **optimal** control frequency?

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Motivations and Problem

- **Problem**: The **control frequency** for a system is a **configurable** environmental parameter.
- Applications
  - Robot control (Kober et al., 2013)
  - Finance, trading (Murphy et al., 2001)

	Control opportunities	Sample complexity
High frequency	✓	✗
Low frequency	✗	✓

- **Research Question**: Can we exploit this **trade-off** to find an **optimal** control frequency?

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

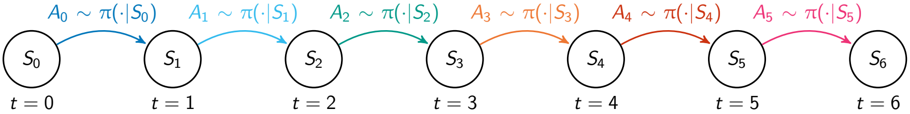
Action Persistence

- **Idea**: **persisting** each action for k consecutive steps

<u>Alberto Maria Metelli</u>, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Action Persistence

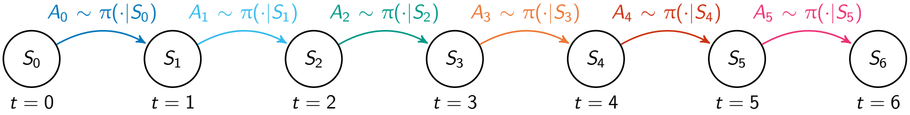
- **Idea**: **persisting** each action for k consecutive steps
- No action persistence



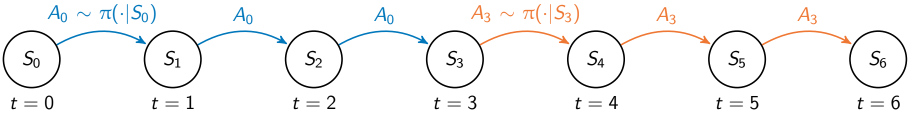
Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Action Persistence

- **Idea**: **persisting** each action for k consecutive steps
- No action persistence



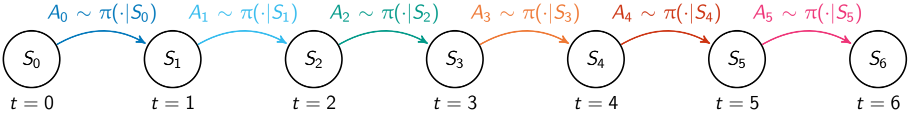
- Action persistence ($k = 3$) $\rightarrow$ **policy view**
  - k-persistent policy (non-Markovian and non-stationary)



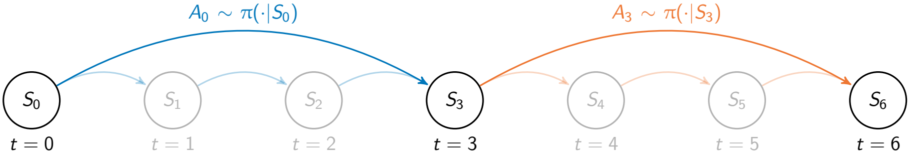
Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Action Persistence

- **Idea**: **persisting** each action for k consecutive steps
- No action persistence



- Action persistence ($k = 3$) $\rightarrow$ **environment view**
  - k-persistent MDP (Conf-MDP)



Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Control Opportunities

- $Q^*_k \leqslant Q^*$ for all $k \geqslant 1$
- How much do we lose by persisting k times the actions?

$$\left\| Q^*_k - Q^* \right\|_{p,\mu} \leqslant \frac{\gamma}{1-\gamma} \frac{1-\gamma^{k-1}}{1-\gamma^k} \left\| \mathcal{W}_1(P^{\pi^*}, P^\delta) \right\|_{p,\mu}$$

- Increasing with k
- $\mathcal{W}_1(P^{\pi^*}, P^\delta)$: Wasserstein distance between transition kernels
  - Can be bounded under Lipschitz conditions (Rachelson and Lagoudakis, 2010)

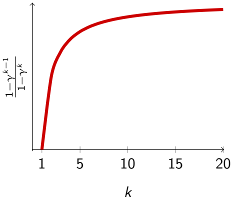
Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Control Opportunities

- $Q^*_k \leqslant Q^*$ for all $k \geqslant 1$
- How much do we lose by persisting k times the actions?

$$\left\| Q^*_k - Q^* \right\|_{p,\mu} \leqslant \frac{\gamma}{1-\gamma} \boxed{\frac{1-\gamma^{k-1}}{1-\gamma^{k}}} \left\| \mathcal{W}_1(P^{\pi^*}, P^{\delta}) \right\|_{p,\mu}$$

- Increasing with k
- $\mathcal{W}_1(P^{\pi^*}, P^{\delta})$: Wasserstein distance between transition kernels
  - Can be bounded under Lipschitz conditions (Rachelson and Lagoudakis, 2010)



Control Opportunities

- $Q^*_k \leqslant Q^*$ for all $k \geqslant 1$
- How much do we lose by persisting k times the actions?

$$\left\| Q^*_k - Q^* \right\|_{p,\mu} \leqslant \frac{\gamma}{1-\gamma} \quad \frac{1-\gamma^{k-1}}{1-\gamma^k} \boxed{\left\| \mathcal{W}_1(P^{\pi^*}, P^{\delta}) \right\|_{p,\mu}}$$

- Increasing with k
- $\mathcal{W}_1(P^{\pi^*}, P^{\delta})$: Wasserstein distance between transition kernels
  - Can be bounded under Lipschitz conditions (Rachelson and Lagoudakis, 2010)

$$P^{\pi^*}(s', a'|s, a) = \pi^*(a'|s') \, P(s'|s, a)$$

$$P^{\delta}(s', a'|s, a) = \delta_a(a') \, P(s'|s, a)$$

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Control Opportunities

- $Q_k^* \leqslant Q^*$ for all $k \geqslant 1$
- How much do we lose by persisting k times the actions?

$$\left\| Q_k^* - Q^* \right\|_{p,\mu} \leqslant \frac{\gamma}{1-\gamma} \quad \frac{1-\gamma^{k-1}}{1-\gamma^k} \quad \left\| \mathcal{W}_1(P^{\pi^*}, P^{\delta}) \right\|_{p,\mu}$$

- Increasing with k
- $\mathcal{W}_1(P^{\pi^*}, P^{\delta})$: Wasserstein distance between transition kernels
  - Can be bounded under **Lipschitz** conditions (Rachelson and Lagoudakis, 2010)

$$\left\| \mathcal{W}_1(P^{\pi^*}, P^{\delta}) \right\|_{p,\mu} \leqslant L_Q \left[(L_{\pi^*} + 1) L_T + \sigma_p \right]$$

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Persistent Fitted Q-Iteration (PFQI)

Fitted Q-Iteration
(Ernst et al., 2005)

- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

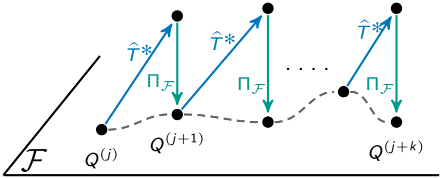
$$Q^{(j+1)} = \Pi_{\mathcal{F}} \widehat{T}^* Q^{(j)}$$

- $Q^{(j)} \rightsquigarrow Q^*$
- What about Q_k^*?

Empirical Bellman Operators

$$(\widehat{T}^* f)(S_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} f(S_{i+1}, a)$$

$$T^* \simeq \Pi_{\mathcal{F}} \widehat{T}^*$$



Persistent Fitted Q-Iteration (PFQI)

Persistent Fitted Q-Iteration

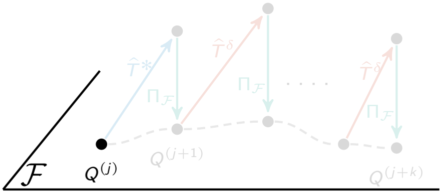
- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

$$Q^{(j+1)} = \begin{cases} \Pi_{\mathcal{F}} \widehat{\mathcal{T}}^{*} Q^{(j)} & \text{if } j \bmod k = 0 \\ \Pi_{\mathcal{F}} \widehat{\mathcal{T}}^{\delta} Q^{(j)} & \text{otherwise} \end{cases}$$

- $Q^{(j)} \rightsquigarrow Q^{*}_{k}$

Empirical Bellman Operators



Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Persistent Fitted Q-Iteration (PFQI)

Persistent Fitted Q-Iteration

- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

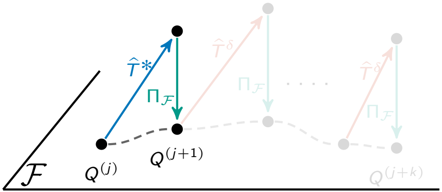
$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

$$Q^{(j+1)} = \begin{cases} \Pi_{\mathcal{F}} \widehat{T}^{*} Q^{(j)} & \text{if } j \bmod k = 0 \\ \Pi_{\mathcal{F}} \widehat{T}^{\delta} Q^{(j)} & \text{otherwise} \end{cases}$$

- $Q^{(j)} \rightsquigarrow Q_k^{*}$

Empirical Bellman Operators

$$(\widehat{T}^{*} f)(S_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} f(S_{i+1}, a)$$



Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Persistent Fitted Q-Iteration (PFQI)

Persistent Fitted Q-Iteration

- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

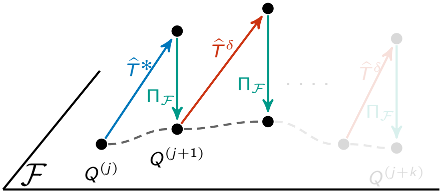
$$Q^{(j+1)} = \begin{cases} \Pi_{\mathcal{F}} \widehat{T}^{*} Q^{(j)} & \text{if } j \bmod k = 0 \\ \Pi_{\mathcal{F}} \widehat{T}^{\delta} Q^{(j)} & \text{otherwise} \end{cases}$$

- $Q^{(j)} \rightsquigarrow Q_k^{*}$

Empirical Bellman Operators

$$(\widehat{T}^{*} f)(S_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} f(S_{i+1}, a)$$
$$(\widehat{T}^{\delta} f)(S_i, A_i) = R_i + \gamma f(S_{i+1}, A_i)$$

$$T_k^{*} = (T^{\delta})^{k-1} T^{*} \simeq (\Pi_{\mathcal{F}} \widehat{T}^{\delta})^{k-1} \Pi_{\mathcal{F}} \widehat{T}^{*}$$



Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Persistent Fitted Q-Iteration (PFQI)

Persistent Fitted Q-Iteration

- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

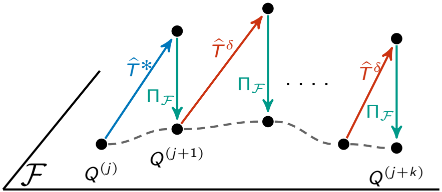
$$Q^{(j+1)} = \begin{cases} \Pi_{\mathcal{F}} \widehat{T}^{*} Q^{(j)} & \text{if } j \bmod k = 0 \\ \Pi_{\mathcal{F}} \widehat{T}^{\delta} Q^{(j)} & \text{otherwise} \end{cases}$$

- $Q^{(j)} \rightsquigarrow Q_k^{*}$

Empirical Bellman Operators

$$(\widehat{T}^{*} f)(S_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} f(S_{i+1}, a)$$
$$(\widehat{T}^{\delta} f)(S_i, A_i) = R_i + \gamma f(S_{i+1}, A_i)$$

$$T_k^{*} = (T^{\delta})^{k-1} T^{*} \simeq (\Pi_{\mathcal{F}} \widehat{T}^{\delta})^{k-1} \Pi_{\mathcal{F}} \widehat{T}^{*}$$



Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Sample Complexity: Error Propagation

$$\left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \frac{2}{1-\gamma} \quad \frac{\gamma^k}{1-\gamma^k} \quad C_k(J,\mu,\nu,p) \quad \mathcal{E}_k(J,\mu,\nu,p)$$

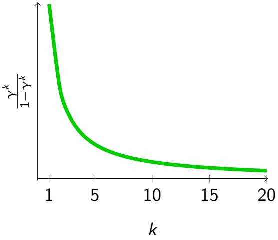
- Decreasing with k
- Concentrability coefficients (Farahmand, 2011)
- Approximation errors → decreasing with number of samples

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Sample Complexity: Error Propagation

$$\left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \frac{2}{1-\gamma} \boxed{\frac{\gamma^k}{1-\gamma^k}} C_k(J,\mu,\nu,p) \quad \mathcal{E}_k(J,\mu,\nu,p)$$

- Decreasing with k
- Concentrability coefficients (Farahmand, 2011)
- Approximation errors → decreasing with number of samples



Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Sample Complexity: Error Propagation

$$\left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \frac{2}{1-\gamma} \frac{\gamma^k}{1-\gamma^k} C_k(J,\mu,\nu,p) \; \mathcal{E}_k(J,\mu,\nu,p)$$

- Decreasing with k
- Concentrability coefficients (Farahmand, 2011)
- Approximation errors → decreasing with number of samples

$$C_k(m) = \sup_{\pi_1,\dots,\pi_a \in \Pi^{\mathrm{SD}}} \left\| \frac{\mathrm{d}\rho (P^\delta)^{k-1} P^{\pi_1} \dots (P^\delta)^{k-1} P^{\pi_a} (P^\delta)^b}{\mathrm{d}\nu} \right\|_{q,\nu} \leqslant \sup_{\pi_1,\dots,\pi_m \in \Pi^{\mathrm{SD}}} \left\| \frac{\mathrm{d}\rho P^{\pi_1} \dots P^{\pi_m}}{\mathrm{d}\nu} \right\|_{q,\nu} = C_1(m)$$

$$a = m \text{ div } k \quad b = m \text{ mod } k$$

<u>Alberto Maria Metelli</u>, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Sample Complexity: Error Propagation

$$\left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \frac{2}{1-\gamma} \quad \frac{\gamma^k}{1-\gamma^k} \quad C_k(J,\mu,\nu,p) \; \boxed{\mathcal{E}_k(J,\mu,\nu,p)}$$

- Decreasing with k
- Concentrability coefficients (Farahmand, 2011)
- Approximation errors → decreasing with number of samples

$$\epsilon^{(j)} = \begin{cases} T^* Q^{(j)} - Q^{(j+1)} & \text{if } j \bmod k = 0 \\ T^\delta Q^{(j)} - Q^{(j+1)} & \text{otherwise} \end{cases}$$

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Control Frequency Trade-Off

$$\left\| Q^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \left\| Q^* - Q_k^* \right\|_{p,\mu} + \left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu}$$

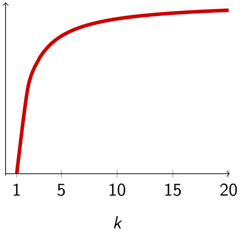
- Control Opportunities
- Algorithm-independent
- **Increasing with k**

- Sample Complexity
- Algorithm-dependent
- **Decreasing with k**

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Control Frequency Trade-Off

$$\left\| Q^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \boxed{\left\| Q^* - Q_k^* \right\|_{p,\mu}} + \left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu}$$



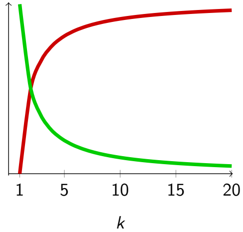
- Control Opportunities
- Algorithm-independent
- **Increasing with k**

- Sample Complexity
- Algorithm-dependent
- **Decreasing with k**

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Control Frequency Trade-Off

$$\left\| Q^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \left\| Q^* - Q_k^* \right\|_{p,\mu} + \left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu}$$



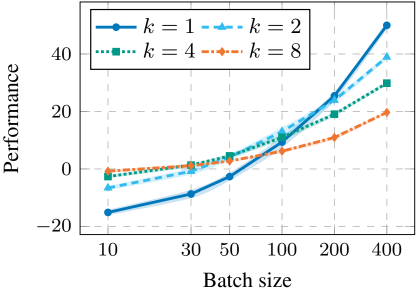
- Control Opportunities
- Algorithm-independent
- **Increasing with k**

- Sample Complexity
- Algorithm-dependent
- **Decreasing with k**

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Forex Trading

- **Task**: USD traded with EUR
- **Positions**: Long, short, flat



Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. *Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning*. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020.

Take-Home Messages

I - Modeling Environment Configurability

Environment configurability emerges in several real-world *scenarios*

II - Learning in cooperative Conf-MDPs

Configuring the environment can improve *agent's optimal performance*

III - Applications of Conf-MDPs

Knowing the agent's policy space *helps environment configuration*

Adapting the control frequency *can improve the learning performance*

Take-Home Messages

I - Modeling Environment Configurability

Environment configurability emerges in several real-world *scenarios*

II - Learning in cooperative Conf-MDPs

Configuring the environment can improve *agent's optimal performance*

III - Applications of Conf-MDPs

Knowing the agent's policy space *helps environment configuration*

Adapting the control frequency *can improve the learning performance*

Take-Home Messages

I - Modeling Environment Configurability

Environment configurability emerges in several real-world *scenarios*

II - Learning in cooperative Conf-MDPs

Configuring the environment can improve *agent's optimal performance*

III - Applications of Conf-MDPs

Knowing the agent's policy space *helps environment configuration*

Adapting the control frequency *can improve the learning performance*

Future Works

I - Modeling Environment Configurability

- **Multiple** agents and **multiple** configurators

II - Learning in Conf-MDPs

- **Online** learning in **cooperative** Conf-MDPs
- Learning in **non-cooperative** Conf-MDPs

III - Applications of Conf-MDPs

- **Online** and **dynamic** action persistence

Future Works

I - Modeling Environment Configurability

- **Multiple** agents and **multiple** configurators

II - Learning in Conf-MDPs

- **Online** learning in **cooperative** Conf-MDPs
- Learning in **non-cooperative** Conf-MDPs

III - Applications of Conf-MDPs

- **Online** and **dynamic** action persistence

Future Works

I - Modeling Environment Configurability

- **Multiple** agents and **multiple** configurators

II - Learning in Conf-MDPs

- **Online** learning in **cooperative** Conf-MDPs
- Learning in **non-cooperative** Conf-MDPs

III - Applications of Conf-MDPs

- **Online** and **dynamic** action persistence

Thank You for Your Attention!

Contact: albertomaria.metelli@polimi.it
Web page: albertometelli.github.io

References I

Dimitri P. Bertsekas. *Dynamic programming and optimal control, 3rd Edition*. Athena Scientific, 2005. ISBN 1886529264.

Bruce Lee Bowerman. Nonstationary markov decision processes and related topics in nonstationary markov chains. 1974.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Kamil Andrzej Ciosek and Shimon Whiteson. OFFER: off-environment reinforcement learning. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1819–1825. AAAI Press, 2017.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.*, 6: 503–556, 2005.

Amir Massoud Farahmand. *Regularization in Reinforcement Learning*. PhD thesis, University of Alberta, 2011.

Víctor Gallego, Roi Naveiro, and David Ríos Insua. Reinforcement learning under threats. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9939–9940. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33019939.

Robert Givan, Sonia M. Leach, and Thomas L. Dean. Bounded parameter markov decision processes. In Sam Steel and Rachid Alami, editors, *Recent Advances in AI Planning, 4th European Conference on Planning, ECP'97, Toulouse, France, September 24-26, 1997, Proceedings*, volume 1348 of *Lecture Notes in Computer Science*, pages 234–246. Springer, 1997. doi: 10.1007/3-540-63912-8_89.

Garud N. Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30(2):257–280, 2005. doi: 10.1287/moor.1040.0129.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In Claude Sammut and Achim G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, pages 267–274. Morgan Kaufmann, 2002.

Jens Kober, J. Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *I. J. Robotics Res.*, 32(11): 1238–1274, 2013. doi: 10.1177/0278364913495721.

References II

Alessandro Lazaric, Marcello Restelli, and Andrea Bonarini. Reinforcement learning in continuous action spaces through sequential monte carlo methods. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 833–840. Curran Associates, Inc., 2007.

Erwan Lecarpentier and Emmanuel Rachelson. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7214–7223, 2019.

Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy poisoning in batch reinforcement learning and control. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 14543–14553, 2019.

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. Configurable markov decision processes. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3488–3497. PMLR, 2018a.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 5447–5459, 2018b.

References III

Alberto Maria Metelli, Emanuele Ghelfi, and Marcello Restelli. Reinforcement learning in configurable continuous environments. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4546–4555. PMLR, 2019a.

Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. Policy space identification in configurable environments. *CoRR*, abs/1909.03984, 2019b.

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. Control frequency adaptation via action persistence in batch reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6862–6873. PMLR, 2020.

Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Arnab Nilim and Laurent El Ghaoui. Robustness in markov decision problems with uncertain transition matrices. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 839–846. MIT Press, 2003.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179, 2018. doi: 10.1561/2300000053.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. AAAI Press, 2010.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 307–315. JMLR.org, 2013.

References IV

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Emmanuel Rachelson and Michail G. Lagoudakis. On the locality of action domination in sequential decision making. In *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2010, Fort Lauderdale, Florida, USA, January 6-8, 2010*, 2010.

Thomas J Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

Jay K. Satia and Roy E. Lave Jr. Markovian decision processes with uncertain transition probabilities. *Oper. Res.*, 21(3): 728–740, 1973. doi: 10.1287/opre.21.3.728.

Rui Silva, Francisco S. Melo, and Manuela Veloso. What if the world were different? gradient-based exploration for new optimal policies. In Daniel D. Lee, Alexander Steen, and Toby Walsh, editors, *GCAI-2018, 4th Global Conference on Artificial Intelligence, Luxembourg, September 18-21, 2018*, volume 55 of *EPiC Series in Computing*, pages 229–242. EasyChair, 2018.

Rui Silva, Gabriele Farina, Francisco S. Melo, and Manuela Veloso. A theoretical and algorithmic analysis of configurable mdps. In J. Benton, Nir Lipovetzky, Eva Onaindia, David E. Smith, and Siddharth Srivastava, editors, *Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling, ICAPS 2018, Berkeley, CA, USA, July 11-15, 2019*, pages 455–463. AAAI Press, 2019.

Saumya Sinha and Archis Ghate. Policy iteration for robust nonstationary markov decision processes. *Optim. Lett.*, 10(8): 1613–1628, 2016. doi: 10.1007/s11590-016-1040-6.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Heinrich Von Stackelberg. *Marktform und gleichgewicht*. J. springer, 1934.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Bernhard Wymann, Eric Espié, Christophe Guionneau, Christos Dimitrakakis, Rémi Coulom, and Andrew Sumner. Torcs, the open racing car simulator. 4:6, 2000.

References V

Haoqi Zhang and David C. Parkes. Value-based policy teaching with active indirect elicitation. In Dieter Fox and Carla P. Gomes, editors, *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 208–214. AAAI Press, 2008.